Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 3781–3790

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

A Survey on RAG with LLMs

Muhammad Arslan[a]*, Hussam Ghanem[a], Saba Munawar[b] and Christophe Cruz[a]

[a]*Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Dijon, France*
[b]*National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan*

Abstract

In the fast-paced realm of digital transformation, businesses are increasingly pressured to innovate and boost efficiency to remain competitive and foster growth. Large Language Models (LLMs) have emerged as game-changers across industries, revolutionizing various sectors by harnessing extensive text data to analyze and generate human-like text. Despite their impressive capabilities, LLMs often encounter challenges when dealing with domain-specific queries, potentially leading to inaccuracies in their outputs. In response, Retrieval-Augmented Generation (RAG) has emerged as a viable solution. By seamlessly integrating external data retrieval into text generation processes, RAG aims to enhance the accuracy and relevance of the generated content. However, existing literature reviews tend to focus primarily on the technological advancements of RAG, overlooking a comprehensive exploration of its applications. This paper seeks to address this gap by providing a thorough review of RAG applications, encompassing both task-specific and discipline-specific studies, while also outlining potential avenues for future research. By shedding light on current RAG research and outlining future directions, this review aims to catalyze further exploration and development in this dynamic field, thereby contributing to ongoing digital transformation efforts.

Keywords: Large Language Models (LLMs); Natural Language Processing (NLP); Retrieval-Augmented Generation (RAG); Text generation; Digital transformation.

1. Introduction

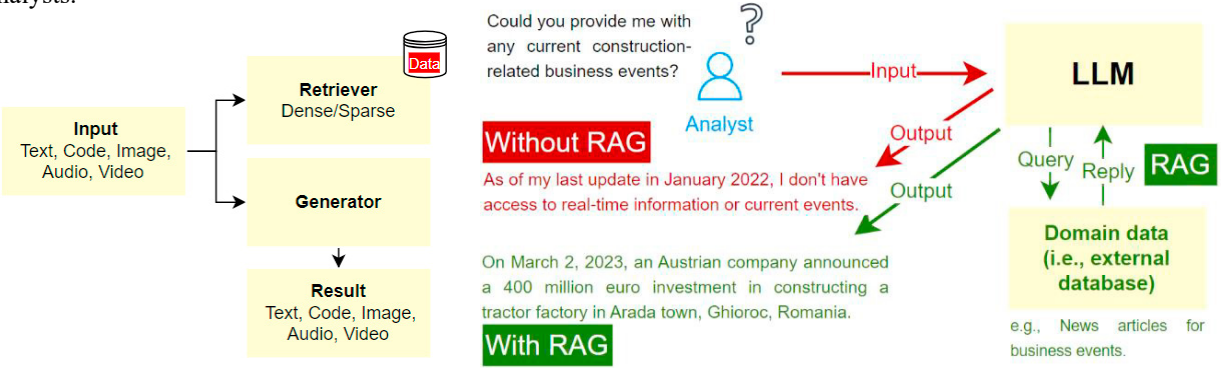
Digital transformation signifies the incorporation of digital technology across different facets of a business, reshaping its operations and value delivery to customers [1]. At the forefront of driving such transformative practices are Large Language Models (LLMs), advanced machine learning models trained extensively on textual data to comprehend and produce human-like text [1]. LLMs, such as the Generative Pre-training Transformer (GPT)

* Corresponding author. Tel.: +33 03 80 39 50 00; fax: +33 03 80 39 50 69.
E-mail address: muhammad.arslan@u-bourgogne.fr

1877-0509
10.1016/j.procs.2024.09.178

series [2, 3] and others, have demonstrated remarkable capabilities in NLP tasks [4]. However, these models face challenges when dealing with domain-specific queries, often generating inaccurate or irrelevant information, commonly referred to as "hallucinations", particularly when data is sparse [5]. This limitation makes deploying LLMs in real-world settings impractical, as the generated output may not be reliable [4].

In the middle of 2020, Lewis et al. [6] introduced RAG, a significant advancement in the field of LLMs for improving generative tasks (see Fig. 1 (a)). RAG incorporates an initial step where LLMs search an external data source to retrieve relevant information before producing text or answering questions. RAG addresses these limitations by integrating external data retrieval into the generative process, thereby enhancing the accuracy and relevance of the generated output. By dynamically retrieving information from knowledge bases during inference, RAG provides a more informed and evidence-based approach to language generation, significantly reducing the risk of hallucinations and improving the overall quality of the generated text [4, 6]. This approach has the potential to make LLMs more practical for real-world applications, as it ensures that the generated output is grounded in retrieved evidence, leading to more reliable and accurate results. Fig. 1 (b) showcases how real-time business systems can leverage the RAG with LLM architecture. As an example, without RAG, the system lacks access to real-time or updated information. However, with RAG integration, leveraging external data sources such as news articles, the system can respond to current business events, presenting opportunities for business intelligence analysts.



(a) (b)

Fig. 1. (a) A generic RAG architecture, where users' queries, potentially in different modalities (e.g., text, code, image, etc.), are inputted into both the retriever and the generator. The retriever scans for relevant data sources in storage, while the generator engages with the retrieval outcomes, ultimately generating results across various modalities [6]; Fig. 1. (b) illustrates how RAG integration with the LLM handles queries that fall outside the scope of the LLM's training data.

While the field of RAG has seen substantial growth, several online surveys [4, 7, 8, 9] have explored technological advancements in RAG. Although these surveys provide valuable insights and references, they offer only a limited overview of RAG applications. To address this gap, this paper aims to provide an exhaustive overview of RAG applications, including both task-specific and discipline-specific studies, as well as future directions. By highlighting the current state of RAG research and its potential future directions, this review aims to inspire further investigation and development in this exciting field.

The paper's structure is as follows: Section 2 presents the adopted research methodology for this survey. In Section 3, we provide an overview of RAG applications, followed by a detailed discussion in Section 4. The paper concludes in Section 5, summarizing the key findings and implications of the study.

2. Background

The research method (see Fig. 2) employed in this paper involves a thorough review and analysis of research publications related to RAG. The main objective is to identify and categorize its applications across various NLP tasks and disciplines. The paper begins by collecting research publications specific to RAG, focusing on their applications. Since the RAG with LLM domain is relatively new and emerging, with many studies available as pre-